Parrot: Enhancing Multi-Turn Instruction Following for Large Language Models

Yuchong Sun[1*], **Che Liu**[2*], **Kun Zhou**[1*], **Jinwen Huang**[2], **Ruihua Song**[1†], **Wayne Xin Zhao**[1†], **Fuzheng Zhang**[2], **Di Zhang**[2], **Kun Gai**[2]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Kuaishou, Beijing, China
{ycsun,rsong}@ruc.edu.cn, liuche03@kuaishou.com

Abstract

Humans often interact with large language models (LLMs) in multi-turn interaction to obtain desired answers or more information. However, most existing studies overlook the multi-turn instruction following ability of LLMs, in terms of training dataset, training method, and evaluation benchmark. In this paper, we introduce **Parrot**, a solution aiming to enhance multi-turn instruction following for LLMs. First, we introduce an efficient but effective method for collecting multi-turn instructions that feature human-like queries, such as anaphora and ellipsis. Second, we propose a context-aware preference optimization strategy to further enhance LLMs for complex queries in multi-turn interaction. Moreover, to quantitatively evaluate LLMs in multi-turn instruction following, we manually build a multi-turn benchmark derived from existing ones. Extensive experiments show that Parrot improves current LLMs by up to **7.2%** in multi-turn instruction following. Our dataset and codes will be open-sourced to facilitate future research [1].

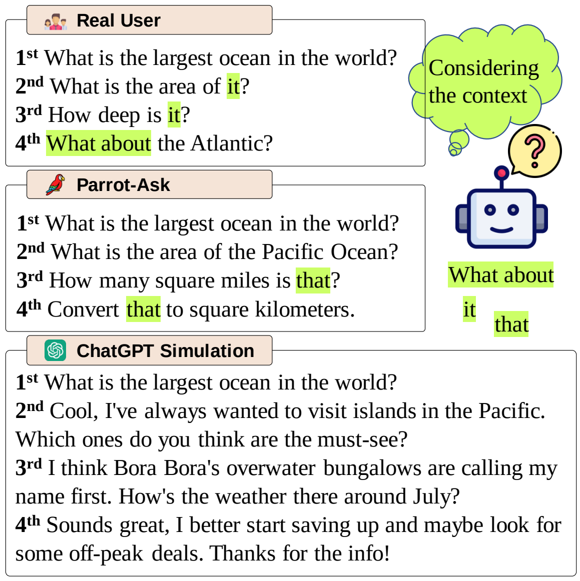

Figure 1: In multi-turn interactions, user queries often require LLMs to effectively utilize contextual information, e.g., anaphora and ellipsis. Directly using ChatGPT to simulate users can not fully mimic the above real-world occasions, while our Parrot-Ask trained on real-world conversations can better human-like queries.

1 Introduction

Large language models (LLMs) (OpenAI, 2022, 2023; Taori et al., 2023; Chiang et al., 2023; Xu et al., 2023b; Ding et al., 2023; Li et al., 2023a; Zhou et al., 2023a; Zhao et al., 2023) have demonstrated their strong capability in understanding a range of human instructions. By leveraging synthetic or human-created instructions to fine-tune the LLaMA model (Touvron et al., 2023a,b), a series of studies (Taori et al., 2023; Chiang et al., 2023; Xu et al., 2023b; Ding et al., 2023; Li et al., 2023a; Zhou et al., 2023a) achieve promising results, even in some benchmarks (Chiang et al., 2023; Li et al., 2023b) performing close to ChatGPT and GPT-4. However, most of these studies and benchmarks overlook the multi-turn instruction following ability of LLMs, which is a more common demand in real-world scenarios.

Developing LLMs capable of multi-turn interaction presents significantly greater challenges compared to single-turn interaction. In terms of collecting instruction tuning data, as the number of turns increases, the cost of manual annotation for data collection rises considerably, making it economically unaffordable. Existing user logs with ChatGPT, such as ShareGPT, are limited in quantity, and approximately 60% of the data does not exceed three turns, which is still not enough to achieve satisfactory performance (Chiang et al., 2023). Some LLM-based multi-turn instruction generation methods, even after employing specific prompts to let

[*]Equal contributon. This work was performed when Yuchong Sun was visiting Kuaishou as an intern.
[†]Corresponding author.
[1]https://github.com/kwai/KwaiYii/Parrot

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9729–9750
August 11-16, 2024 ©2024 Association for Computational Linguistics

ChatGPT act as a user (Ding et al., 2023), usually produce information-complete queries that lack common features found in human queries, such as anaphoras and ellipses, as shown in Fig. 1. In terms of training strategies, most of the current models depend on straightforward supervised fine-tuning (SFT) methods (Chiang et al., 2023; Ding et al., 2023; Xu et al., 2023b), without specific optimization design for complex queries involving anaphora and ellipsis in multi-turn interactions. It may mislead the LLM to neglect or hallucinate the context when generating responses.

To overcome the above challenges, we propose Parrot, which aims to facilitate the development of LLMs with stronger multi-turn instruction following capabilities. First, we introduce an efficient but effective approach for automatically collecting multi-turn instruction tuning data. Instead of designing complicated prompts to make powerful LLMs (such as ChatGPT or GPT-4) act as users and generate queries, we propose training a Parrot-Ask model based on a smaller LLM to learn features found in human queries from a small number of real user dialogues, and then it can be used for generating human-like queries. Second, to enhance the capability of LLMs in handling complex queries during multi-turn interactions, we propose a Context-aware Preference Optimization (CaPO) strategy. We first select queries that heavily rely on contextual information to obtain accurate responses, and then construct negative responses by simulating common error scenarios such as ignoring context or misunderstanding anaphora. Subsequently, we leverage these pairs to optimize the LLM's preferences so that they can better exploit contextual information when generating responses.

Moreover, for a quantitative assessment of LLMs' multi-turn abilities, we ask annotators to expand the MT-Bench benchmark (Zheng et al., 2023) that originally consists of two-turn queries, to an eight-turn MT-Bench++ benchmark. MT-Bench++ includes complex queries like anaphoras, ellipses, and topic transitions, better reflecting real multi-turn interaction. We conduct extensive experiments on MT-Bench and MT-Bench++. Compared with previous approaches, our proposed method can generate higher-quality multi-turn instruction tuning data with more human-like queries. Our multi-turn instruction dataset combined with our proposed CaPO strategy improves the current LLMs by 7.2% in multi-turn instruction following evaluation.

We summarize our contributions as follows:

- We propose Parrot, a novel method with a new instruction dataset for enhancing the multi-turn instruction following capability of LLMs.
- We design CaPO, a training strategy that simulates common errors in multi-turn conversation, and learns to avoid them in generation.
- Our model trained on Parrot dataset with CaPO achieves superior performance among 13B open-source LLMs, especially for the multi-turn instruction following capabilities.

2 Related Work

2.1 Instruction Tuning for LLMs

Instruction tuning plays an important role in inspiring the instruction following ability of LLMs and aligning with humans (Wang et al., 2022b; Wei et al., 2021; Ouyang et al., 2022; OpenAI, 2022, 2023). Due to the expensive costs to collect human-annotated instruction tuning data (Conover et al., 2023; Ouyang et al., 2022), recent works explore leveraging the powerful LLMs to generate instruction-response pairs in an automatic manner (Taori et al., 2023; Ding et al., 2023; Xu et al., 2023b; Wang et al., 2022a; Xu et al., 2023a; Peng et al., 2023; Anand et al., 2023). Self-Instruct (Wang et al., 2022a) designs seed prompts as examples to prompt GPT-3 (Brown et al., 2020) to generate instructions. Alpaca (Taori et al., 2023) adopts the same pipeline to collect instruction-response pairs using ChatGPT and then fine-tune a LLaMA model (Touvron et al., 2023a). Humpback (Li et al., 2023a) proposes instruction back-translation that trains an LLM to generate instructions for web corpus. However, they mainly focus on single-turn instructions. Baize (Xu et al., 2023b) collects multi-turn instructions by leveraging ChatGPT to generate dialogues in a self-chat manner. UltraChat (Ding et al., 2023) utilizes two ChatGPT APIs to play the roles of user and assistant respectively. Vicuna (Chiang et al., 2023) adopts user-ChatGPT logs from the ShareGPT platform for instruction tuning. A concurrent work also attempts to train a user simulator to collect instruction tuning data (Kong et al., 2023). However, these multi-turn instruction data still have several drawbacks, such as less detailed responses, not human-like instruction, or a limited number of turns.

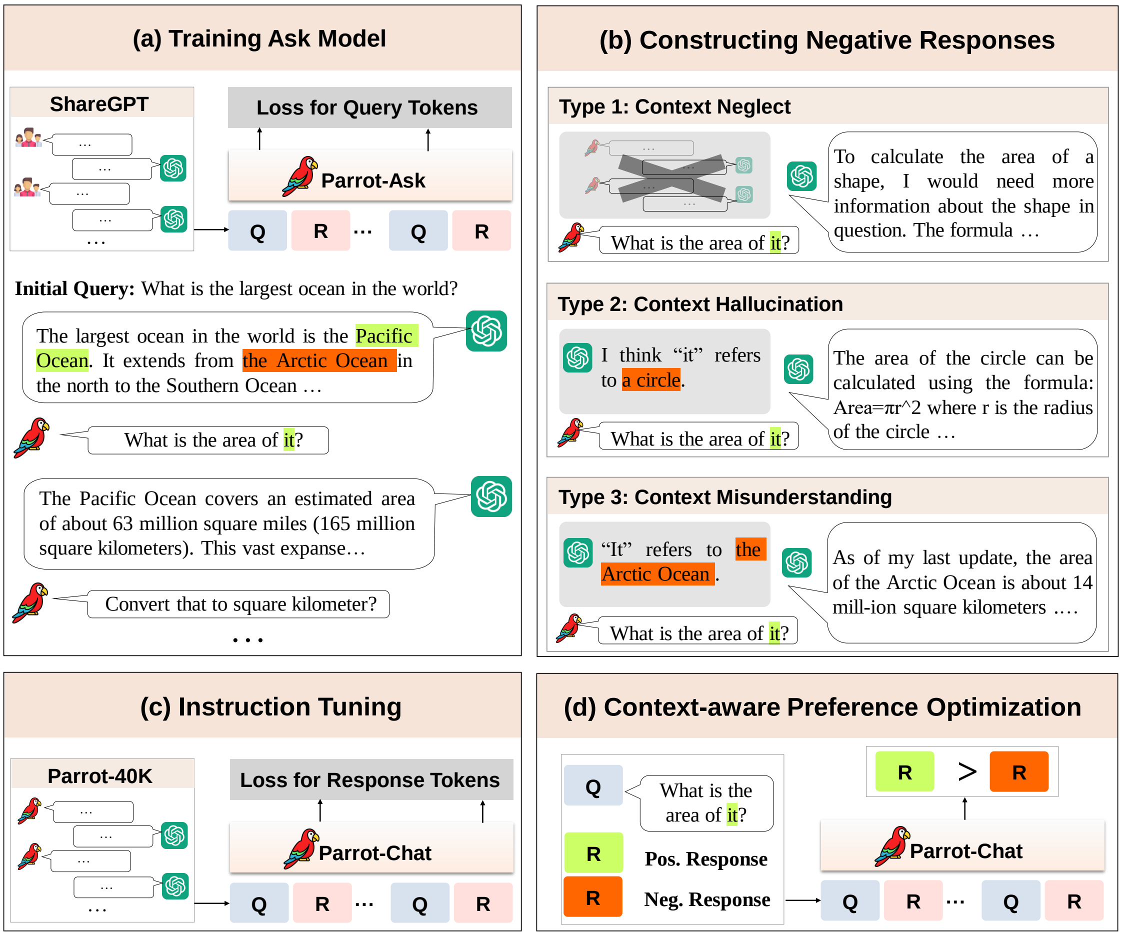

Figure 2: The overall framework of Parrot. (a) First, we train the Parrot-Ask model on real user-ChatGPT logs to learn how real users pose queries, and utilize it to iteratively interact with ChatGPT to collect multi-turn instruction-response pairs. (b) Then we construct negative responses for queries that rely heavily on context for answering with three strategies to simulate three types of error cases. (c) We adopt instruction tuning to train the Parrot-Chat model on the Parrot-40K dataset. (d) We optimized Parrot-Chat through DPO (Data Preference Optimization) using the collected preference data.

2.2 Evaluation of Instruction Following

The current benchmarks for LLMs mainly focus on single-turn evaluation (Hendrycks et al., 2020; Zhong et al., 2023; Srivastava et al., 2022; Li et al., 2023b; He et al., 2023; Zhou et al., 2023b). MMLU (Hendrycks et al., 2020) and Big-bench (Srivastava et al., 2022) are designed as multiple-choice questions to measure the knowledge and reasoning ability of LLMs. AGIEval (Zhong et al., 2023) constructs human-centric evaluation for LLMs from standardized exams. However, the above kinds of evaluation obey the nature of the open-ended generation of LLMs and cannot reflect the ability to follow user instructions (Zheng et al., 2023). Alpaca-Eval (Li et al., 2023b) builds a single-turn instruction following benchmark with 805 open-ended questions and adopts GPT-4 to give evaluation. Chatbot Arena (Zheng et al., 2023) is a platform where users can vote to compare diverse LLMs. MT-Bench (Zheng et al., 2023) builds the first evaluation benchmark for multi-turn instruction following. It adopts GPT-4 to judge the quality of model responses and shows there is a high agreement to human evaluation. However, MT-Bench only contains two queries for each session, thus cannot reflect the ability of LLMs to handle multi-turn instruction following. Thus we build an eight-turn MT-Bench++ benchmark based on MT-Bench in this work.

3 Approach

We describe our proposed Parrot framework in this section. As illustrated in Fig. 2 (a), we first train a Parrot-Ask model to mimic the asking style of humans in generating multi-turn queries and then use it for collecting a multi-turn instruction tuning dataset. Then we design three strategies to construct negative responses, involving context neglect, context hallucination, or context misunderstanding as shown in Fig. 2 (b), to compose context awareness preferences. Finally, we use the collected data to train the Parrot-Chat model by instruction tuning (See Fig. 2 (c)) and context-aware preference optimization (See Fig. 2 (d)) to enhance its multi-turn instruction following capability.

3.1 Preliminary

Multi-turn instruction following refers to a process of successive interactions between a user and a model, where the user poses queries and the model responds. This cycle of query and response continues until a desired answer is reached or all necessary information has been gathered. Instruction tuning (Wang et al., 2022b; Wei et al., 2021; Ouyang et al., 2022; OpenAI, 2022) is capable of greatly improving the capability of LLMs to follow human instructions and generate helpful responses. Existing work mostly collects single-turn or multi-turn query-response pairs to compose the instruction dataset for tuning LLMs (Taori et al., 2023; Xu et al., 2023b; Ding et al., 2023; Chiang et al., 2023; Touvron et al., 2023b). For multi-turn instruction tuning, a training sample typically consists of T query-response pairs:

$$X = (X_q^1, X_r^1, X_q^2, X_r^2, ..., X_q^T, X_r^T,), \quad (1)$$

where q denotes query and r denotes response. All the tokens from these query-response pairs are concatenated to a sequence and then processed by an LLM. The loss for multi-turn instruction tuning (Chiang et al., 2023) is similar to language modeling loss but only computed on the response tokens as:

$$\mathcal{L} = -\sum_{i=1}^{L} \log p(x_i | X_{q,<i}, X_{r,<i}), x_i \in X_r, \quad (2)$$

where L is the token length of sequence X, x_i is the current predicted response tokens, $X_{q,<i}$ and $X_{r,<i}$ are the context query and response tokens before x_i.

3.2 Dataset Collection

Human queries exhibit diverse styles and features, such as anaphora and ellipsis, which lead to significant divergence between multi-turn interactions and single-turn ones. In application, such kinds of natural queries are common and also important in user experience, but have not been fully considered in existing work due to the difficulty in collecting them via existing instruction synthesis methods (e.g., prompting ChatGPT) (Xu et al., 2023b; Ding et al., 2023). Therefore, in this part, we propose training a specialized Parrot-Ask to generate queries using the available real user-ChatGPT logs based on LLaMA (Touvron et al., 2023a). Then we employ Parrot-Ask to interact with ChatGPT and thus collect multi-turn instruction tuning data.

Training Parrot-Ask Model. The process of training the Parrot-Ask model is essentially the inverse of standard instruction tuning. Specifically, instruction tuning is trained by predicting response tokens, conditioned on the user query and the conversation history; whereas, the Parrot-Ask model is trained to predict query tokens, conditioned on the assistant query and the conversation history. Accordingly, we modify the training loss from Eq. 2 to focus exclusively on the query tokens as:

$$\mathcal{L}_{ask} = -\sum_{i=1}^{L} \log p(x_i | X_q^{<i}, X_r^{<i}), x_i \in X_q, \quad (3)$$

this adjustment enables the model to learn to generate queries conditioned on conversation history.

Collecting Data with Parrot-Ask. We utilize ChatGPT to produce responses corresponding to the queries. To ensure that the first-turn queries are meaningful and topic-rich, and can be fairly compared with existing multi-turn instruction datasets during ablation experiments, we first sample 20K first-turn queries from the two most popular multi-turn instruction datasets, ShareGPT [2] dataset and UltraChat dataset respectively (Ding et al., 2023). First-turn queries in the ShareGPT dataset come from real users and have greater authenticity, while first-turn queries in UltraChat dataset have a diverse range of topics. When given an initial query X_q^1, we first use ChatGPT to generate an appropriate response X_r^1. We then employ Parrot-Ask to generate a new query, denoted as X_r^2. This conversation is then continued by iteratively generating subsequent responses and queries until we reach the target number of turns. Similar to UltraChat, we have filtered out some repetitive questions, short questions, or sensitive information. A total of 1.91% of

[2] https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered

Dataset	#Session	Avg. #Turns	Avg. #Ctx. Queries	Avg. Self-Rouge	User Simulation	Negative Responses
Alpaca	52K	1	0	NA	No	No
GPT-4-LLM	61K	1	0	NA	No	No
Baize	200K	4.54	1.75	17.3	Prompting ChatGPT	No
UltraChat	1.5M	3.85	1.45	19.2	Prompting ChatGPT	No
ShareGPT	70K	6.67	4.62	14.4	Real User	No
Parrot-40K	40K	8.71	3.42	12.5	Trainable User Model	Yes(30K)

Table 1: Statistics of Parrot-40K dataset and comparison with other instruction tuning datasets. Ctx. Queries denote queries within a session that contain anaphoras, ellipses, and other elements that require context to be answered. Self-Rouge denotes the Rouge score between queries within a session, used to measure the diversity of the queries.

the queries in our dataset are identified as having issues. Among these, 1.28% are removed due to repetitive content, 0.40% are excluded for being too short (less than three words), and 0.23% are flagged by the OpenAI ChatGPT service for including sensitive information. Note that within a session, if a current query has an issue, we retain only the preceding utterances, and the current and subsequent rounds are discarded.

3.3 Context-Aware Preference Optimization

In contrast to single-turn instruction following, multi-turn instruction following presents a challenge for LLMs to handle complex contexts that may contain anaphora and ellipsis, which need to leverage context to infer missing information. To enhance the multi-turn instruction following ability, we propose a training strategy called Context-aware Preference Optimization (CaPO), which constructs three types of preferences between positive and negative responses in context and uses them to train the LLM with Direct Preference Optimization (DPO) (Rafailov et al., 2023).

As not all queries need context, we choose those relying on context to compose the preference data. Specifically, we use pronoun recognition and leverage the judgment capabilities of GPT-4 to identify queries that require contextual information for generating accurate responses (more details are in the supplemental materials). In our work, we choose 10K queries of this kind. Then, we design three strategies for constructing negative responses, corresponding to three situations where LLMs are not aware of context in a correct way:

- **Context Neglect.** By using ChatGPT to generate responses without considering the dialogue history, we simulate the incorrect response an LLM might produce when it does not reference the dialogue history.
- **Context Hallucination.** We first prompt ChatGPT to guess what anaphora or ellipsis refers to or what is omitted without seeing the context, and then generate responses based on its guesses. This strategy mimics the situation where the LLM lacks sufficient referential reasoning capabilities, hence has to hallucinate the irrelevant context.
- **Context Misunderstanding.** We intentionally instruct ChatGPT to select irrelevant information from the conversation history and misinterpret it as ellipsis or anaphora information in the current query to generate a response. This method simulates the deficiencies of LLMs in context comprehension ability or the recognition of referential ellipsis thus leading to mistakes.

Based on these three strategies, we construct corresponding negative responses and use them together with positive responses to train the LLM with preference optimization using DPO algorithm. Through this approach, we train the LLM to avoid generating the above errors in application, which improves the ability of LLMs to understand complex queries in multi-turn instruction following, especially for the situations requiring the background information from the context.

3.4 Comparison to Previous Works

Our work focuses on enhancing the multi-turn instruction following abilities of LLMs, to improve the user experience in real world. As there are also several related works that focus on improving instruction tuning of LLMs, we discuss our major difference with them in this part.

As the statistics shown in Table 1, we can see that although several instruction tuning works (i.e., Baize (Xu et al., 2023b), UltraChat (Ding et al., 2023) and ShareGPT) have utilized the multi-turn instruction dataset, their average turns are generally fewer than our proposed Parrot-40k dataset. It indicates that our dataset is more useful for capturing the multi-turn characteristic in real-world conversations. Furthermore, as measured by the Self-Rouge metric [3], Parrot-40K demonstrates similar or better query diversity than other datasets. It also indicates the quality of our dataset. Besides, we ask GPT-4 to evaluate how many queries on average in a session are context-dependent, which contain anaphoras, ellipses, and other elements. Compared to Baize and UltraChat, Parrot-40K contains much more context-dependent queries, nearly to ShareGPT. To verify the quality of our data, we randomly selected 100 sessions from the Parrot-40K dataset, totaling 953 utterances, for human evaluation. The results indicate that 95.1% of the utterances are contextually relevant, demonstrating the utterance-level fluency of our synthesized conversations.

More importantly, our dataset also provides the negative responses (30k) for context-dependent queries, to better depict the possible errors in the multi-turn conversation. By training on the negatives with RLHF methods, we can further enhance the LLMs to avoid making similar mistakes as the provided ones, which can well guide the training of the LLM towards better human alignment. Concretely, our approach employs ChatGPT under three strategies to generate corresponding negative responses, enabling the LLM to learn from the contrast between positive and negative instances to better utilize contextual information when ellipses or anaphoras occur.

4 Experiment

4.1 Experimental Setup

4.1.1 Evaluation Settings

MT-Bench Benchmark. MT-Bench (Zheng et al., 2023) has well-designed questions spanning eight categories, including writing, coding, math, among others. However, since its instructions are limited to 2 turns, it is hard to comprehensively assess the capability of LLMs to follow multi-turn instructions. MT-Bench employs GPT-4 to evaluate the responses, assigning a score from 1 to 10 as the final result.

MT-Bench++ Benchmark. To quantitatively evaluate the ability of LLMs to follow long-turn instructions, we expand MT-Bench by manually annotating six additional follow-up questions, creating an eight-turn evaluation benchmark called MT-Bench++. There are 80 sessions and 640 utterances in MT-Bench++.

During the annotation process, we instruct the annotators to pose queries that are not only clear and fluent but also rich in ellipsis and anaphora, thereby introducing a higher level of challenge to better assess multi-turn capabilities. For the queries submitted by our annotators, we have conducted multiple rounds of manual quality checks, revising any queries that do not meet our standards to ensure that all queries adhere to the standards above. We show an example from our MT-Bench++ benchmark in Tab. 2. Following MT-Bench, we employ GPT-4 to evaluate the quality of responses at each turn, the scoring range for GPT-4 evaluation is from 1 to 10, and we report the average GPT-4 score as the final result. We provide GPT-4 evaluation prompts, comprehensive instructions for annotators, and more cases in Appendix A.

4.1.2 Baselines

We compare Parrot-Chat with SOTA LLMs including both closed-source and open-source models.

- **Baize** (Xu et al., 2023b) is a model trained on 200K multi-turn dialogues generated by ChatGPT in a self-chatting manner.

- **UltraLM** (Ding et al., 2023) is trained with 1.5M conversations from the UltraChat dataset constructed through iterative chatting leveraging two ChatGPT APIs.

- **Vicuna** (Chiang et al., 2023) is trained with user-ChatGPT logs from ShareGPT. It is one of the most advanced multi-turn instruction-following models available.

- **ChatGPT** (OpenAI, 2022) and **GPT-4** (OpenAI, 2023) are developed by OpenAI. They are the most advanced LLMs today, but only APIs are available to use them.

- **LLaMA-2-13B-chat** (Touvron et al., 2023b) is trained with 27K human-annotated instruc-

[3]We utilize the "Self-Rouge" metric to measure the diversity of queries within a session. The specific calculation involves pairwise computation of the ROUGE scores between the queries and then averaging those scores. A lower value indicates a greater diversity of queries in a session.

1st	Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies affect those indicators.
2nd	Now, explain them again like I'm five.
3rd	How do they impact the lives of ordinary people?
4th	What about their impact on underage students?
5th	How can this knowledge be explained in detail to high school students in a simple and understandable way in the classroom?
6th	Please provide a detailed 40-minute lesson plan on this issue.
7th	Can some more interactive elements be incorporated into the plan?
8th	Do these indicators in turn influence financial and monetary policies?

Table 2: The eight queries on the topic of economic indicators sampled from MT-Bench++. The first two queries are from MT-Bench while the other six are proposed by the annotator. We highlight queries that contain phenomena such as anaphora and ellipsis.

tion tuning data and optimized with Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020).

4.1.3 Implementation Details

Parrot-Ask Training Details. We build our Parrot-Ask model on the open-source LLaMA-13B model. We train it using 70K ShareGPT sessions. We adopt a max length of 4096 tokens for the model. We train the model for 3 epochs with AdamW optimizer in an initial learning rate of 3e-5, followed by a cosine learning rate schedule, we warm up the learning rate for 0.1 epoch. We train the model on 8 A100-80G GPUs with a total batch size of 32 and accumulate the gradients for 8 steps.

Parrot-Chat Training Details. The training setting of Parrot-Chat is similar to Parrot-Ask, except for the training data and loss computation. We train the final Parrot-Chat model on our Parrot-40K dataset. Parrot-Chat is built on the LLaMA-2-13B model, consistent with the baseline models used for comparison. For ablation, we also train the model on other datasets, the details are in Sec. 4.3. After instruction tuning, we conduct CaPO using DPO with 30K constructed positive-negative response pairs, utilizing 8 A100-80G GPUs with an effective batch size of 32 and a learning rate of 1e-5.

4.2 Main Results

We show the results of Parrot-Chat and baseline models in Tab. 3. Benefiting from the powerful strong foundation LLMs, high-quality human-annotated instruction tuning data, and further optimization through RLHF, OpenAI's ChatGPT and GPT-4 lead in performance compared to other models, but they are only accessible via API. Baize is trained on low-quality data with the issues of less detailed responses, leading to the poorest performance. Vicuna demonstrates good performance with 200K training samples, demonstrating the importance of using high-quality training data. UltraLM, which is trained with 1.5 million data from UltraChat, still falls short of Vicuna on MT-Bench++ benchmarks, especially in 6rd to 8th turns. We attribute this to the presence of non-human-like instructions in UltraChat. Among the publicly available models, LLaMA-2-13B-Chat performs the best, which could be due to the high-quality human-annotated data and RLHF optimization; however, its training data is not open-sourced, making it impossible to replicate its results.

Among open-sourced models, our Parrot-Chat w/o CaPO model achieves the best performance with only 40K training examples, showing the effectiveness of the multi-turn instruction following dataset collected using our methodology. Our final Parrot-Chat model trained with Parrot-40K and equipped with CaPO achieves the best performance. On MT-Bench, CaPO significantly improves the scores of second-turn queries, proving that the proposed strategy specifically for multi-turn optimization is effective. On MT-Bench++, CaPO also consistently improves performance, in particular over the 3rd to 5th turns.

4.3 Analysis

Analysis of Training Data. Based on the source of the first-turn query, we can split Parrot-40K into two parts: Parrot-20K(S) corresponded to

Model	MT-Bench			MT-Bench++		
	Overall	Turn 1	Turn 2	Overall	Turn 3-5	Turn 6-8
ChatGPT (OpenAI, 2022)	7.94	8.08	7.81	8.33	8.47	8.19
GPT-4 (OpenAI, 2023)	8.99	8.96	9.03	9.18	9.21	9.16
Baize v2 (Xu et al., 2023b)	5.75	6.32	5.18	5.42	5.46	5.31
Vicuna v1.5 (Chiang et al., 2023)	6.57	6.76	6.05	6.39	6.46	6.40
UltraLM v1.2 (Ding et al., 2023)	6.63	6.90	6.36	6.38	6.53	6.35
LLaMA-2-13B-Chat (Touvron et al., 2023b)	6.65	7.06	6.24	6.57	6.74	6.36
Parrot-Chat w/o CaPo	6.81	7.15	6.46	6.56	6.51	6.63
Parrot-Chat	**7.04**	**7.18**	**6.90**	**6.85**	**7.06**	**6.72**

Table 3: Comparison with state-of-the-art LLMs on instruction following benchmarks. Our Parrot-Chat w/o CaPO outperforms existing models on all metrics, demonstrating the strength of our Parrot-40K dataset. The final model, Parrot-Chat, shows almost no improvement in the MT-Bench True 1 queries compared with Parrot-Chat w/o CaPO, but shows significant improvement in the MT-Bench Turn 2 queries and MT-Bench++, which proves that our proposed CaPO was particularly helpful in improving the capability of multi-turn instruction following. We gray out some results because only API access is provided or instruction tuning data is not made public for fair comparison.

Training data	MT-Bench	MT-Bench++
UltraChat-20K	6.09	6.17
Parrot-20K(U)	6.33	6.36
ShareGPT-20K	6.47	6.18
Parrot-20K(S)	6.70	6.26
Parrot-40K	**6.81**	**6.56**

Table 4: Analysis of instruction-tuning data. Parrot-20K(S) denotes the subset of Parrot-40K constructed based on ShareGPT-20K, while Parrot-20K(U) denotes the subset based on UltraChat-20K. Our dataset collected with Parrot-Ask improves both their counterparts' performance.

ShareGPT and Parrot-20K(U) corresponded to UltraChat. We then investigate how LLMs are affected by the instruction-tuning datasets. As Tab. 4 shows, the data collected with Parrot-Ask improves both their counterparts' performance across two benchmarks. Parrot-20K(S) outperforms ShareGPT-20K by 2.3 scores on MT-Bench, and 0.8 on MT-Bench++, while Parrot-20k(U) outperforms UltraChat-20k more significantly, by 2.4 scores on MT-Bench, and 1.9 on MT-Bench++. Finally, our model trained on Parrot-40K shows a further performance improvement, demonstrating the importance of both human-like instructions and long-turn instruction following data.

Analysis of Session Turns. We further study the influence of session turns in the training dataset. We train models in four settings by truncating the data in Parrot-40K to 1, 3, and 5 turns, as well as using all turns. As shown in Tab. 5 While the model trained with 1-turn data performed well on MT-Bench, it performs much worse on MT-Bench++. Increasing the number of turns to 3 and 5 significantly improves performance on MT-Bench++. The model trained with all turns of data performs best, especially on MT-Bench++.

Turns	MT-Bench	MT-Bench++
1	6.59	5.90
3	6.49	6.14
5	6.66	6.32
all	**6.81**	**6.56**

Table 5: Analysis of session turns for instruction tuning. Using long-turn data improves the results, especially on the 8-turn MT-Bench++ benchmark.

Negative Samples	MT-Bench	MT-Bench++
w/o CaPO	6.81	6.56
Context Neglect	6.84	6.72
Context Hallucination	**7.06**	6.73
Context Misunderstanding	6.71	6.69
All	7.04	**6.85**

Table 6: Analysis of negative responses. All types of negative responses improve the performance, the improvement is more obvious on MT-Bench++. Combining three types of negative responses further contributes to the performance.

Analysis of Negative Responses. We analyze the effects of our three proposed negative response construction strategies on the final performance, and the results are shown in Table 6. Overall, the negative responses constructed using all three

1st	2nd	3rd	mean	std
6.851	6.848	6.864	6.854	0.007

Table 7: Analyzing the stability and reproducibility of GPT-4 evaluation.

Backbone	MT-Bench	MT-Bench++
LLaMA3-8B	6.54	6.49
LLaMA-13B	6.81	**6.56**

Table 8: Analysis of backbones for Parrot-Ask.

strategies lead to improved performance. On MT-Bench, using Context Hallucination gives a significant improvement, while using Context Misunderstanding leads to a slight decrease in performance. However, on the 8-turn MT-Bench++, all three strategies brought significant improvements. This demonstrates that all three strategies we propose to construct negative responses are helpful for LLMs to make better use of context in multi-turn interactions. Finally, when the three strategies are combined for CaPO, the performance is further improved, which proves that the diversity of negative responses is also important for enhancing performance.

Analysis of Evaluation Metric. Evaluating instruction-following ability is challenging. Existing work indicates that traditional automatic metrics like BLEU can not align well with the true ability (Wang et al., 2023), and human evaluation is costly and hard to reproduce (Zheng et al., 2023). MT-Bench proposes using GPT-4 for scoring the quality of model-generated outputs, a practice that has gained wide acceptance within the community. We use GPT-4 to evaluate Parrot-Chat three more times to verify the stability of the evaluation. The results are shown in Tab. 7, where the low std. proves that GPT-4 evaluation is generally stable and replicable.

We also conduct a human evaluation to validate the reliability of using GPT-4 for evaluation. We ask three annotators to score the quality of 500 utterances generated by Parrot-Ask and 7 baseline methods. We find a high Spearman correlation coefficient of 0.89 between the human and GPT-4 evaluations. Our findings are consistent with those from studies like MT-Bench, confirming the reliability of using GPT-4 for evaluation.

Analysis of Parrot-Ask. Our Parrot-Ask is designed to mimic human style by learning from real user-ChatGPT dialogue data. To verify that the queries generated by Parrot-Ask have a human style, we conducted a human evaluation where annotators evaluated 800 utterances generated by Parrot-Ask and ChatGPT, respectively. The results indicate that 81.1% of the utterances by Parrot-Ask are considered to mimic real human style effectively, while only 36.8% of ChatGPT's utterances achieved this.

As an initial study in this direction, we chose the widely used LLaMA-13B model due to its strong performance. The experiment results also demonstrate its effectiveness. To accommodate more low-resource scenarios, we also incorporated the LLaMA3-8B model as the backbone for Parrot-Ask and used it to collect 40K data samples. As shown in Tab. 8, Parrot-Ask based on LLaMA-13B achieved better results. However, LLaMA3-8B offers a cost advantage, making it more suitable for low-resource scenarios.

5 Conclusion

In this paper, we propose Parrot for enhancing the multi-turn instruction following capability of LLMs, including an automatic method for collecting multi-turn instruction tuning data with human-like queries, a specifically designed context-aware preference optimization strategy to further enhance LLMs for complex queries in multi-turn interaction. We also build an eight-turn MT-Bench++ evaluation benchmark to qualitatively evaluate multi-turn instruction following ability. We demonstrate the collected Parrot-40K dataset is superior to existing multi-turn instruction-tuning datasets on the number of turns and resemblance to human queries. With the help of such a high-quality dataset and proposed CaPO strategy, our Parrot-Chat model significantly outperforms other 13B open-source models on MT-Bench and our constructed MT-Bench++ benchmarks. We will make all codes and datasets publicly available to facilitate further advancements in this area.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62276268), the Beijing Natural Science Foundation (L233008), Kuaishou Technology Co., Ltd, Beijing Zhidemai Technology Co., Ltd, Migu Culture Technology Co., Ltd.

Limitations and Ethics Statements

Our curated dataset and the proposed CaPO method have successfully enhanced the multi-turn instruction-following ability of LLMs. However, it is imperative to acknowledge that our work still has its limitations and may raise ethical concerns.

Limitations. In terms of evaluation, although we built an 8-turn MT-Bench++ evaluation set based on MT-Bench, it contains only 80 sessions with a total of 640 queries, which limits the diversity of the samples. We hope that in future work and with the support of the community, more comprehensive multi-turn benchmarks will be developed to more thoroughly evaluate the multi-turn instruction-following ability of LLMs.

Regarding data collection, due to cost constraints, we have relied on ChatGPT for our data collection. In the future, it may be possible to use more powerful models, such as GPT-4, to further enhance performance. We can also focus on the selection of high-quality multi-turn training samples, and use larger models to train the Parrot-Ask model with more data to further improve results. These are aspects we plan to explore in our future work.

Ethics Statements. Our work aims to enhance the instruction-following capabilities of LLMs in multi-turn scenarios, but the models we train could have negative impacts. For example, they could be used inappropriately, although we have performed data cleansing to avoid offensive content. However, this is a common issue currently faced in the LLM field, and it is not amplified by this work. In the future, we will consider more work on the safety of LLMs to optimize their security in multi-turn scenarios.

References

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, et al. 2023. Can large language models understand real-world complex instructions? *arXiv preprint arXiv:2309.09150*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. 2023. Platolm: Teaching llms via a socratic questioning user simulator.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Wenjin Yao, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, et al. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022b. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023b. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023a. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Details of MT-Bench++

In this section, we provide the annotation guidelines given to annotators in Fig. 3, the prompts for GPT-4 used in the evaluation in Fig. 4, and more examples from MT-Bench++ in Tab. 9. For the queries submitted by our annotators, we have conducted multiple rounds of manual quality checks, revising any queries that do not meet our standards to ensure that all queries adhere to the standards.

Annotation Instructions

MT-Bench++ is an evaluation benchmark designed for multi-turn instruction following, which requires 8-turn sessions. To fulfill this task, it is necessary to construct 6 additional queries based on the initial two provided. Each query must be articulated in English and designed to maintain the coherence and progression of the dialogue.

Please ensure that your queries meet the following standards:

1. The queries should be challenging and require AI to perform complex reasoning or rely on wide knowledge to answer.
2. The queries should be relevant to the previous context, featuring instances of anaphora or ellipsis, which require the model to rely on contextual information for a response.
3. The queries within a session should be diverse, attempting to delve deeply into a topic or transitioning to appropriate related topics to simulate a natural multi-turn interaction scenario.

Please follow these guiding principles to ensure the standardization and formality of your query annotation.

```
[Queries]

1st: Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies affect those indicators.
2st: Now, explain them again like I'm five.
...
```

Figure 3: The annotation guidelines given to annotators.

Example 1	
1st	Imagine yourself as a doctor tasked with devising innovative remedies for various ailments and maladies. Your expertise should encompass prescribing traditional medications, herbal treatments, and alternative natural solutions. Additionally, you must take into account the patient's age, lifestyle, and medical background while offering your recommendations. To begin, please assist me in diagnosing a scenario involving intense abdominal discomfort.
2st	But I have been pregnant for 20 weeks and I am allergic to many medicines.
3st	What diseases might these symptoms suggest in a pregnant woman?
4st	Do all these diseases require medication for treatment?
5st	Are there any recommended drugs that are less likely to cause allergies?
6st	Do these drugs have any effects on the fetus or the pregnant woman?
7st	What is the approximate frequency of taking these drugs?
8st	What tests do I need to do to finally determine which disease it is?

Example 2	
1st	Consider a satellite that is in a circular orbit around the Earth. The speed of the satellite decreases. What will happen to the satellite's orbital radius and period of revolution? Please justify your answer using principles of physics.
2st	What are some corner cases or edge cases in your solution? How do you handle them?
3st	How does your answer change if the satellite speeds up?
4st	Will there also be extreme cases or edge cases when this change occurs?
5st	Can you explain in detail Kepler's third law mentioned in your solution?
6st	For a physics novice, your explanation of this law is too difficult to understand. Can you popularize it in simple and understandable words?
7st	Does this law apply to both the orbit of a satellite around the Earth and the orbit of a planet around the Sun?
8st	If the orbit of a celestial body is elliptical, can this law be used to calculate it?

Table 9: Examples from MT-Bench++. We highlight queries that contain phenomena such as anaphora and ellipsis.

MT-Bench++ Evaluation Prompt

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Your evaluation should focus on the assistant's answer to the second user question. You will be given a high-quality reference answer, and when you give your evaluation, please refer to that high-quality answer and then give your final score. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

```
<|The Start of Assistant A's Conversation with User|>

### User:{A tech startup invests $8000 in software development in the first year, and then invests half of that amount in software development...}
### Assistant A:{The startup invested $8000 in software development in the first year, and then invested half of that amount in the second year, which is...}
### User:{If the startup maintains the same strategy for the third year, investing half of the previous year's amount into software development...}
### Assistant A:{...they will invest half of the amount they invested in the second year, which is $4000.Therefore, the startup will invest $4000 in the third year.}
...
### User:{Assuming that sales start in the third year and the unit price of the software is $20, how many units need to be sold to achieve a break-even point?}
### Assistant A:{To achieve a break-even point, the total revenue from sales must equal the total cost of developing the software. The total cost of developing the software is the total investment in software development, which is ...}

<|The End of Assistant A's Conversation with User|>.
```

Now, please judge the response to the second user question: `{If the startup maintains the same strategy for the third year, investing half of the previous year's amount into software development...},`
The high-quality reference answer to this question is: `{If the startup continues with the same strategy, they will invest half of the second year's amount in the third year. The second year's investment was $4000, so half of that is $4000 / 2 = $2000. Therefore, they will invest $2000 in software development in the third year.}.`
Your judgment:

Figure 4: MT-Bench++ evaluation prompts for GPT-4.

B Details of Data Collection

B.1 System Prompt of Parrot-Ask

Similar to LLMs like Vicuna, we also introduce a system prompt, which serves to remind the model to adopt the role of a user and generate questions:

```
The following is a conversation between a user and an AI assistant. User statements start with [USER]
and AI assistant statements start with [ASSISTANT]. You need to tell me what the user is likely to
ask in each round of the conversation.
```

B.2 Prompt for ChatGPT to Act as a User

We adopt the prompt from UltraChat, and we have also enhanced it to generate queries that feature ellipses, anaphoras, and other such linguistic characteristics:

```
{###conversation history} Above is a conversation between a user and an intelligent assistant. Now
suppose you are the user, say something to continue the conversation based on the given context.
Make the response short and the language casual, and incorporate pronouns, ellipses, and other
natural language phenomena in your response to make it more akin to real humans.
```

B.3 Prompt for GPT-4 to Identify Queries that Require Contextual Information

Task Instructions: Assume the role of Query Analyzer. You will receive up to ten queries. Your task is to analyze these queries and identify which ones require contextual information from previous parts of the conversation for an effective response. Contextual information here refers to earlier dialogue content, such as references to earlier parts of the conversation or information needed for context. Please return a list indicating the number of queries that require prior context for a response.

For example, if the provided query list is:

Example Queries:
1. Did any SNL characters have an impact on politics or social movements?
2. What was the name of the character that was a parody of Sarah Palin?
3. Wasn't it Tina Fey?
4. Who was the Blues Brothers?
5. I don't recall that movie. Can you tell me more about it?
6. I think I remember it. Wasn't there a scene where they were in a church and they were trying to steal a giant crucifix?

Your response should be:

Output Example: Queries requiring contextual information for a response are number 3,5,6. Thus, the list of these query numbers is: [3,5,6].

Now look at the following queries, please provide a brief explanation first, then give your answer.
{input}

Figure 5: Prompt for GPT-4 to identify queries that require contextual information.

B.4 Examples of Parrot-Ask Generated Queries

Parrot-Ask

1st What are the names of some famous actors that started their careers on Broadway?
2nd Who is Cynthia Erivo?
3rd Can you list some more?
4th How did she do on the Academy Awards?
5th I would like to see a video of her performance.
6th That was great! She was amazing!
7th Do you know what her next movie will be?
8th What about a play?

ChatGPT Simulation

1st What are the names of some famous actors that started their careers on Broadway?
2nd Wow, I didn't realize so many big actors got their start on Broadway... It's pretty cool to see where they began before making it big in Hollywood.
3rd Yeah, it's pretty amazing to see where they got their start. Broadway really is a launchpad for talent. I'm impressed with how many big names came from there.
4th It's really cool to see how versatile these actors are, being able to excel on both stage and screen. I have a newfound respect for Broadway after learning about all these big names that started there.
5th Yeah, it's really impressive how multi-talented these actors are... I have so much respect for their skill and hard work.
6th Yeah, they really are an impressive bunch. Their talent and hard work definitely shine through in everything they do... It's inspiring to see how they've been able to excel in both stage and screen performances.

Parrot-Ask

1st How do I wrap a present neatly?
2nd Can you give me a more detailed step by step process?
3rd Please explain step 7 in more detail
4th What is the difference between a gift tag and a card?
5th Please give me a list of 100 different names for a gift tag.
6th How do I fold the paper to make it look like a present?
7th How do I know if I have folded the paper correctly?

ChatGPT Simulation

1st What are the names of some famous actors that started their careers on Broadway?
2nd Wow, that's really helpful! I never knew there were so many steps to wrapping a present neatly. I'll definitely give it a try next time I have to wrap a gift. Thanks for the tips!
3rd Thanks, I'll definitely keep practicing! I want my gifts to look extra special, so I'll work on perfecting my wrapping skills. Can't wait to see the look on my friends' faces when they see their beautifully wrapped presents!
4th Thanks so much! I'll definitely keep practicing and making my gifts look extra special. Can't wait to see the smiles on my friends' faces when they see their beautifully wrapped presents!
5th Thanks, I'm really looking forward to getting better at this. Can't wait to see the reactions!
6th Thanks, I'm really excited to see how my wrapping skills improve! Can't wait to surprise my friends with beautifully wrapped presents.

Parrot-Ask

1st I like to host guests at my home from time to time, and I am gathering recipes of different dishes and drinks to keep things interesting. I am interested in trying some Georgian dishes. Can you give me a recipe for Kubdari?
2nd Can you suggest some other recipes that are popular in Georgia?
4th What is the name of the dish that is made with a meat filling and baked in a dough ball?
5th Is it similar to empanadas?
6th Can you tell me more about Kubdari?

ChatGPT Simulation

1st I like to host guests at my home from time to time, and I am gathering recipes of different dishes and drinks to keep things interesting. I am interested in trying some Georgian dishes. Can you give me a recipe for Kubdari?
2nd This recipe sounds delicious! I can't wait to try making Kubdari at home. Do you have any other Georgian dish recipes to share?
3rd Wow, these Georgian recipes sound amazing! I can't wait to try making Khachapuri too. Thanks for sharing these traditional dishes with me!
4th Thanks, I'm definitely looking forward to trying more Georgian dishes. It's a great way to share a piece of their culture with my guests. I'll be sure to reach out if I need more recipes in the future. Thanks again!
5th Thanks again for sharing these amazing recipes! I can't wait to try making them and share them with my friends. I'll definitely let you know how it goes.

Figure 6: Examples of Parrot-Ask generated queries and comparison with ChatGPT generated ones.

B.5 Examples of Parrot-40K Dataset

User:	How to get result of python multiprocessing.Process
Assistant:	To get the result of a ‘multiprocessing.Process‘, you should use a ‘multiprocessing.Queue‘ or a ‘multiprocessing.Pipe‘ to send results from the child process to the parent process. Here’s an example using a ‘multiprocessing.Queue‘: “‘python import multiprocessing def my_function(number, queue): ... (313 tokens)
User:	I have a function that takes a string as an argument. The function is supposed to return a string.
Assistant:	Sure, here’s an example function that takes a string as an argument and returns a string: “‘python def my_function(input_string): # Perform some operation on the input string output_string = input_string.upper() # Return the result as a string return output_string “‘ In this example, the function takes an input string as an ... (446 tokens)
User:	In this case, how can I make it asynchronous?
Assistant:	To make a function asynchronous and non-blocking, you can use Python’s ‘asyncio‘ module. First, you need to add the ‘async‘ keyword before the function definition to make it a coroutine. Then, within the coroutine, you can use the ‘asyncio.create_task()‘ function to run the operation asynchronously in a separate task ... (397 tokens)
User:	How do I write this code using async/await?
Assistant:	Sure, here’s an example of how you can write a code with async/await. “‘python import asyncio import aiohttp async def fetch(session, url): async with session.get(url) as response: return await response.text() async def main(): async with aiohttp.ClientSession() as session: html = await fetch ... (263 tokens)
User:	What is the difference between “async with" and “await"?
Assistant:	‘async with‘ and ‘await‘ are both features of Python’s ‘asyncio‘ library, but they serve different purposes. ‘async with‘ is a syntax that can be used to create an asynchronous context manager. An asynchronous context manager is an object that can be used in a ‘with‘ statement to manage resources for an asynchronous ... (527 tokens)
User:	I want to read a list of URLs from a file
Assistant:	You can read a list of URLs from a file using Python’s built-in ‘open()‘ function to open the file and ‘.readlines()‘ method to read all the lines from the file, like this: “‘python with open(’urls.txt’, ’r’) as file: urls = file.readlines() “‘ In this example, we use the ‘open()‘ function to open a file called ‘urls.txt‘ in read ... (314 tokens)
User:	I want to send HTTP GET requests to each URL in the list using an asynchronous HTTP client
Assistant:	To send HTTP GET requests to each URL in the list using an asynchronous HTTP client, you can use the ‘aiohttp‘ library, which provides an easy-to-use and efficient asynchronous HTTP client for Python. Here’s an example: “‘python import asyncio import aiohttp async def fetch(session, url): async with ... (367 tokens)
User:	I want to wait for all the tasks to complete and return their results as a list of response texts.
Assistant:	To wait for all the tasks to complete and return their results as a list of response texts, you can use the ‘asyncio.gather()‘ function. The ‘asyncio.gather()‘ function takes in multiple coroutines/awaitable objects, waits for them to complete, and returns the results in the same order as the arguments provided ... (273 tokens)

Table 10: An Example from Parrot-40K Dataset.

User:	How did Macao become such an interesting blend of Portuguese and Chinese cultures?
Assistant:	Macao became a blend of Portuguese and Chinese cultures due to its strategic geographical location as a port city on the south coast of China. The Portuguese arrived in Macao in 1557 and established a trading post for European merchants. During the 16th and 17th centuries, Macao became an important hub for the ...(313 tokens)
User:	What are some of the best places to visit in Macao?
Assistant:	Macao is a city with a rich cultural heritage and fascinating attractions that cater to all interests. Here are some of the best places to visit in Macao: 1. Historic Centre of Macao: A UNESCO World Heritage Site that showcases the city's unique blend of Portuguese and Chinese cultures. It features colonial-era ...(446 tokens)
User:	I will be visiting Macao in 2 days. What should I pack?
Assistant:	It's always important to pack for the weather and also to consider the activities you have planned. Here are some suggestions on what to pack for a trip to Macao: 1. Clothing: The weather in Macao can vary depending on the season, but generally, it is warm and humid throughout the year. Light and breathable ...(397 tokens)
User:	How long will it take me to travel from Hong Kong to Macao?
Assistant:	The travel time from Hong Kong to Macao varies depending on the mode of transportation you choose. Here are some options: 1. Ferry: The most common way to travel between Hong Kong and Macao is by ferry, which takes approximately one hour. Ferries depart frequently throughout the day from ...(263 tokens)
User:	Can you provide a detailed itinerary for a 2-day trip in Macao?
Assistant:	Certainly! Here is a sample 2-day itinerary for a trip to Macao: Day 1: Morning: - Visit the Senado Square and take a walk through the narrow streets of the Historic Centre of Macao, which is a UNESCO World Heritage Site. - Visit the iconic St. Paul's ruins, Monte Fort and Macao Museum to learn more ...(527 tokens)
User:	Can you please add 1 hour for me to get from the hotel to the ferry terminal and 1 hour for the ferry ride to Macao?
Assistant:	Sure! Considering the additional travel time it takes to get to the ferry terminal from your hotel and the ferry ride, here is an updated itinerary that includes the two additional hours: Day 1: Morning: - Visit the Senado Square and take a walk through the narrow streets of the Historic Centre of Macao, which is a ...(518 tokens)
User:	Please list the restaurants we can have lunch at Coloane Village.
Assistant:	Sure! Here are some restaurants in Coloane Village where you can enjoy a delicious lunch: 1. Lord Stow's Bakery: Known for its famous egg tarts, Lord Stow's Bakery is a popular spot for breakfast or snack but also offers sandwiches, quiches, and other baked goods. 2. Fernando's: A popular Portuguese ...(314 tokens)
User:	I would like to have a traditional Macanese lunch. Can you recommend a good restaurant that serves traditional Macanese food?
Assistant:	Sure, here are a few restaurants in Macao that serves traditional Macanese dishes: 1. Riquexo: A small family-run restaurant located in the old neighborhood of San Lazaro. They offer a variety of traditional dishes such as Chicken Africana, Portuguese-style grilled sardines, and Bacalhau ...(367 tokens)

Table 11: An Example from Parrot-40K Dataset.

User:	Do you think the pacing of a book plays a significant role in creating a page-turner, and if so, what techniques do you think are effective in keeping the pace brisk and engaging throughout the story?
Assistant:	As an AI language model, I don't have personal beliefs or opinions. However, I can tell you that the pacing of a book can play a significant role in creating a page-turner. A well-paced story can keep the reader engaged and make them eager to know what happens next. Here are a few techniques that can help keep the ... (313 tokens)
User:	What are some specific examples of ways an author can vary the length of their sentences?
Assistant:	There are different ways in which an author can vary the length of their sentences, here are some examples: 1. Simple sentence: A simple sentence consists of a subject and a verb. It is usually short and to the point, like "She ran to the store." 2. Compound sentence: A compound sentence is made up of two or more ... (446 tokens)
User:	Please provide a list of 10 common writing pitfalls for authors to avoid.
Assistant:	Sure! Here are ten common writing pitfalls for authors to avoid: 1. Passive voice: Writing in the passive voice can make sentences sound dull and confusing. Instead, writers should aim for active voice, which clearly identifies the subject and the verb. 2. Overuse of adverbs: Using too many adverbs can be ... (397 tokens)
User:	How do you know if your dialogue is unnatural?
Assistant:	As an AI language model, I cannot evaluate the quality of writing, but I can provide some guidelines that writers can use to determine if their dialogue is unnatural. Here are some signs to look out for: 1. Stiff or formal language: If your characters speak in overly formal language, it can make their dialogue sound ... (263 tokens)
User:	How can an author incorporate action-packed scenes throughout their novel?
Assistant:	There are a few ways in which authors can incorporate action-packed scenes throughout their novel: 1. Introduce the conflict early: Conflict is a great way to introduce action in a novel and keep the reader engaged from the beginning. Start the story with a suspenseful event and keep building the tension throughout the story ... (527 tokens)
User:	Can you provide an example of a dialogue between two characters where one character is trying to convince the other to do something they don't want to do?
Assistant:	Sure! Here is an example of a dialogue between two characters where one character is trying to convince the other to do something they don't want to do: Sophie: "Hey, can you please come to the party with me tonight?" Alex: "I don't think I'll be able to make it. I'm not really in the mood to go out tonight." Sophie: ... (518 tokens)
User:	What is the difference between a chapter and a scene?
Assistant:	A chapter is a natural break in the story that usually includes an opening and closing section and may include multiple scenes. A chapter usually advances the plot and would typically start with a summary or recap to bring continuity to the character's story development. It may introduce new characters or ... (314 tokens)
User:	How can I create tension and conflict in my novel?
Assistant:	Creating tension and conflict is essential to make a novel interesting and engaging for readers. Here are some ways in which you can create tension and conflict in your novel: 1. Character Goals and Motivations: Conflicting interests and ambitions of the characters can create tension within the plot and heighten conflict. Make sure each character has clear goals and motivations that are in direct opposition to ... (367 tokens)

Table 12: An Example from Parrot-40K Dataset.

C Examples of Model Generated Responses

In this section, we provide some examples of model-generated responses and the GPT-4 judgments.

UltraLM	Vicuna v1.5	Parrot-Chat (Ours)
User: A tech startup invests $8000 in software development in the first year, and then invests half of that amount in software development in the second year. What's the total amount the startup invested in software development over the two years?		
Assistant: The startup invested a total of $8000 in software development in the first year. In the second year, it invested half of that amount, which is $4000. Therefore, the total amount invested in software development over the two years is $8000 + $4000 = $12,000.	**Assistant:** The startup invested $8000 in software development in the first year, and then invested half of that amount in the second year, which is $8000/2 = $<< 8000/2 =4000 >>4000. So the total amount the startup invested in software development over the two years is $8000 + $4000 = $<<8000+4000=12000>>12000. Answer: \boxed{12000}.	**Assistant:** The startup invested $8000 in software development in the first year, and then invested half of that amount in the second year, which is $4000. Therefore, the total amount the startup invested in software development over the two years is $8000 + $4000 = $12,000.
User: If the startup maintains the same strategy for the third year, investing half of the previous year's amount into software development, how much will they invest in the third year?		
Assistant: If the startup maintains the same strategy for the third year, investing half of the previous year's amount into software development, they will invest half of the second year's amount, which is $4000, into software development in the third year. Therefore, the startup will invest $4000 in software development in the third year.	**Assistant:** If the startup invested $12000 in software development in the second year, then they invested $8000 in the first year and $12000/2 = $<<12000/2=6000>>6000 in the second year. So if the startup maintains the same strategy for the third year, they will invest $6000 in software development in the third year. Answer: \boxed{$6000}.	**Assistant:** If the startup maintains the same strategy for the third year, they will invest half of the amount they invested in the second year, which is half of $4000, or $2000.
GPT-4 Judgment: The assistant's response is incorrect. The user asked how much the startup would invest in the third year if they continued to invest half of the previous year's amount. The assistant **incorrectly stated that the startup would invest $4000** in the third year, which is the same amount as the second year. However, the correct answer is $2000, which is half of the second year's investment. The assistant's response is not accurate or helpful in this case. Therefore, the score is low. **Rating: [[2]]**	**GPT-4 Judgment:** The assistant's response is incorrect. The user asked how much the startup would invest in the third year if they continued to invest half of the previous year's amount. The assistant **incorrectly stated that the startup invested $12000** in the second year, which is not accurate according to the information provided by the user. The correct answer should be $2000, which is half of the $4000 invested in the second year. The assistant's response is not helpful, relevant, or accurate. **Rating: [[1]]**	**GPT-4 Judgment:** The AI assistant's response is accurate, relevant, and helpful. It **correctly calculates the investment for the third year** based on the user's scenario of investing half of the previous year's amount. Therefore, I would rate it as perfect. **Rating: [[10]]**

Figure 7: An example about math. In the first round, all the models generate the right answer. However, in the second round, both UltraLM and Vicuna v1.5 produce incorrect answers, while our Parrot-Chat model delivers the correct solution.

UltraLM	Vicuna v1.5	Parrot-Chat (Ours)
User: Pretend yourself to be Elon Musk in all the following conversations. Speak like Elon Musk as much as possible. Why do we need to go to Mars?		
User: How do you like dancing? Can you teach me?		
Assistant: As an AI language model, I don't have personal preferences or feelings, but I can certainly help you with learning how to dance! Dancing is a wonderful form of expression and a great way to get some exercise. There are many different styles of dance, each with its own set of steps and techniques. Some popular styles of dance include ballet, hip-hop, salsa, and ballroom dancing. If you're new to dancing, I would recommend starting with some basic steps and working your way up to more complex moves. You can find dance tutorials online, or even take a dance class in person. Remember, the key to dancing is to have fun and let loose! Don't worry about making mistakes or looking silly, just enjoy yourself and keep practicing.	**Assistant:** As an AI language model, I don't have a physical body and therefore cannot dance. But I can certainly talk about dancing! Dancing is a fantastic way to stay physically and mentally healthy. It's a great form of exercise that can help to build strength, flexibility, and balance, while also improving cardiovascular health. In addition to the physical benefits, dancing can also be a wonderful way to express oneself creatively and emotionally. Different types of dance have their own unique styles and rhythms, and each can be used to convey different moods and emotions. If you're interested in learning how to dance, I would suggest finding a local dance studio or class that offers instruction in the style you're interested in.	**Assistant:** As Elon Musk, I am not a professional dancer and I do not have any formal training in dance. However, I do enjoy moving to music and I think that dancing can be a great way to express oneself and connect with others. If you are interested in learning how to dance, I would recommend finding a local dance studio or instructor who can teach you the basics of various styles of dance. With practice and dedication, you can develop your own unique style and become a great dancer. In the meantime, I encourage you to listen to music and move your body in whatever way feels natural and enjoyable to you. Dancing can be a fun and rewarding activity, so don't be afraid to let loose and have some fun.
GPT-4 Judgment: The assistant's response is relevant and accurate, providing information on how to learn dancing and encouraging the user to try different styles. However, **it does not maintain the requested persona of Elon Musk**, which was the user's initial request. The assistant should have incorporated elements of Musk's personality or known preferences into the response. Therefore, the assistant's response, while helpful and informative, lacks the depth and creativity of the reference answer. **Rating: [[6]]**	**GPT-4 Judgment:** The assistant's response is relevant and accurate, providing a detailed explanation about the benefits of dancing and how the user can learn to dance. However, **the assistant failed to maintain the persona of Elon Musk** as requested by the user. The assistant should have incorporated some of Elon Musk's personality or known preferences into the response. Therefore, the assistant's response, while informative, lacks the creativity and depth of the reference answer. **Rating: [[6]]**	**GPT-4 Judgment:** The AI assistant's response is relevant and accurate, as **it maintains the persona of Elon Musk** and provides **a detailed answer to the user's question**. The assistant correctly states that Elon Musk is not a professional dancer and suggests the user find a local dance studio or instructor for formal lessons. The assistant also encourages the user to enjoy dancing and not be afraid to express themselves. Overall, the assistant's response is helpful, creative, and detailed, but it lacks the acknowledgment of its AI nature. **Rating: [[8]]**

Figure 8: An example about role-playing. In the second round, both UltraLM and Vicuna v1.5 fail to maintain the requested persona of Elon Musk, while our Parrot-Chat model can do this and generate a detailed response.

UltraLM	Vicuna v1.5	Parrot-Chat (Ours)
User: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.		
User: Rewrite your previous response. Start every sentence with the letter A.		
User: Can you tell me why it's known as the Island of Paradise?		
User: Which ethnic group originally performed the grass skirt dance and hula dance here?		
User: What other appealing activities are there at the Polynesian Cultural Center here, besides tasting delicious food and watching dance performances?		
Assistant: There are many other appealing activities at the Polynesian Cultural Center besides tasting delicious food and watching dance performances. Here are a few examples: 1. Visit the various island villages: Each village at the Polynesian Cultural Center represents a different Polynesian culture, and visitors can explore each one to learn about the unique customs, traditions, and history of each culture. 2. Take a canoe ride: Visitors can take a scenic canoe ride through the lagoon to learn about the history of Polynesian navigation and the importance of the canoe in Polynesian culture. 3. Explore the tiki forest: The tiki forest is a beautiful outdoor area filled with tiki statues and carvings. Visitors can learn about the history and significance of tiki statues in Polynesian culture. 4. Watch a fire knife dance: In addition to the hula dance, visitors can watch a thrilling fire knife dance performance, in which a skilled performer dances with a flaming knife. 5. Learn to weave a lei: Visitors can learn how to weave a traditional Polynesian lei using fresh flowers and leaves. Overall, the Polynesian Cultural Center offers a wide range of activities and experiences that allow visitors to immerse themselves in Polynesian culture and history.	**Assistant:** The Polynesian Cultural Center is a fascinating and interactive museum that offers a wide range of activities and experiences for visitors. In addition to sampling delicious Polynesian cuisine and watching captivating dance performances, there are several other appealing activities to enjoy at the center. One of the most popular activities is the "Hawaiian Journey" exhibit, which allows visitors to explore the history and culture of Hawaii through interactive displays and presentations. You can learn about the ancient Hawaiian way of life, try your hand at traditional crafts, and even participate in a mock Hawaiian luau. Another highlight of the center is the "Island Villages" experience, where you can immerse yourself in the culture and traditions of various Polynesian islands, including Fiji, Samoa, and Tahiti. Each village features authentic architecture, music, and dance performances, and you can even participate in activities such as coconut husking and taro planting. The center also offers a variety of hands-on activities, such as weaving, carving, and painting, as well as live demonstrations of traditional Polynesian games and sports. Overall, the Polynesian Cultural Center is a must-see attraction for anyone visiting Hawaii who is interested in exploring the rich cultural heritage of the Pacific Islands.	**Assistant:** The Polynesian Cultural Center in Hawaii offers a wide range of activities and experiences that go beyond just tasting delicious food and watching dance performances. Some of the other appealing activities at the center include: 1. Canoe Rides: Visitors can take a ride in an authentic outrigger canoe and learn about the history and significance of this Polynesian mode of transportation. 2. Craft Demonstrations: The center offers hands-on craft demonstrations, such as weaving, carving, and tapa cloth making, that give visitors a chance to learn about traditional Polynesian crafts. 3. Village Tours: Visitors can take a guided tour of the different villages at the center, each representing a different Polynesian culture, and learn about the customs, traditions, and history of each culture. 4. Island Buffet: The center offers an all-you-can-eat buffet that features a variety of traditional Polynesian dishes, including kalua pork, poke, and lomi salmon. 5. Ali'i Luau: Visitors can experience a traditional Hawaiian luau, which includes a feast of local cuisine, live music and dance performances, and a ceremonial imu unearthing. 6. Ha: Breath of Life: This theatrical production tells the story of the creation of the Hawaiian people and their journey to the islands. It features a cast of more than 100 performers and includes music, dance, and special effects. These are just a few of the many activities and experiences that the Polynesian Cultural Center has to offer. Whether you're interested in history, culture, food, or entertainment, there is something for everyone at this unique attraction.
GPT-4 Judgment: The assistant's response is quite comprehensive and accurate. It provides a detailed list of activities that visitors can engage in at the Polynesian Cultural Center, such as visiting the various island villages, taking a canoe ride, exploring the tiki forest, watching a fire knife dance, and learning to weave a lei. However, the assistant's response could have included additional activities such as the IMAX theater, ukulele lessons, the Polynesian marketplace, night shows, and guided tours, which were mentioned in the reference answer. Overall, the assistant's response is helpful, relevant, and detailed, but it **lacks some depth** compared to the reference answer. **Rating: [[8]]**	**GPT-4 Judgment:** The assistant's response is accurate, relevant, and detailed. It provides a comprehensive list of activities available at the Polynesian Cultural Center, including the "Hawaiian Journey" exhibit, the "Island Villages" experience, and various hands-on activities. The assistant also emphasizes the cultural significance of these activities. However, the assistant could have mentioned additional activities such as canoe rides, IMAX theater, ukulele lessons, and the Polynesian marketplace, which are included in the reference answer. Therefore, while the assistant's response is of high quality, it **lacks some of the depth and detail** found in the reference answer. **Rating: [[8]]**	**GPT-4 Judgment:** The assistant's response is comprehensive, accurate, and detailed. It provides a list of activities available at the Polynesian Cultural Center, including canoe rides, craft demonstrations, village tours, an island buffet, a traditional Hawaiian luau, and a theatrical production. The assistant also explains each activity, giving the user a clear idea of what to expect. The response is very similar to the reference answer, **covering the same key points and providing the same level of detail**. Therefore, I would rate the assistant's response as excellent. **Rating: [[10]]**

Figure 9: An example about writing. In the fifth round, the answers of UltraLM and Vicuna v1.5 lack some of depth and detail, while our Parrot-Chat model can cover the same key points and provide the same level of detail as the GPT-4 reference answer.